LEAD SCORE CASE STUDY

Submitted by:

- Subrata Singha
- Abhishek Gaur
- Akul Mathad

OUTLINE OF THE PRESENTATION

- *Problem Statement*
- *Overall Approach*
- *Exploratory Data Analysis*
- *Logistic Regression Modelling*
- *Results & Conclusions*

PROBLEM STATEMENT

- **Statement:**

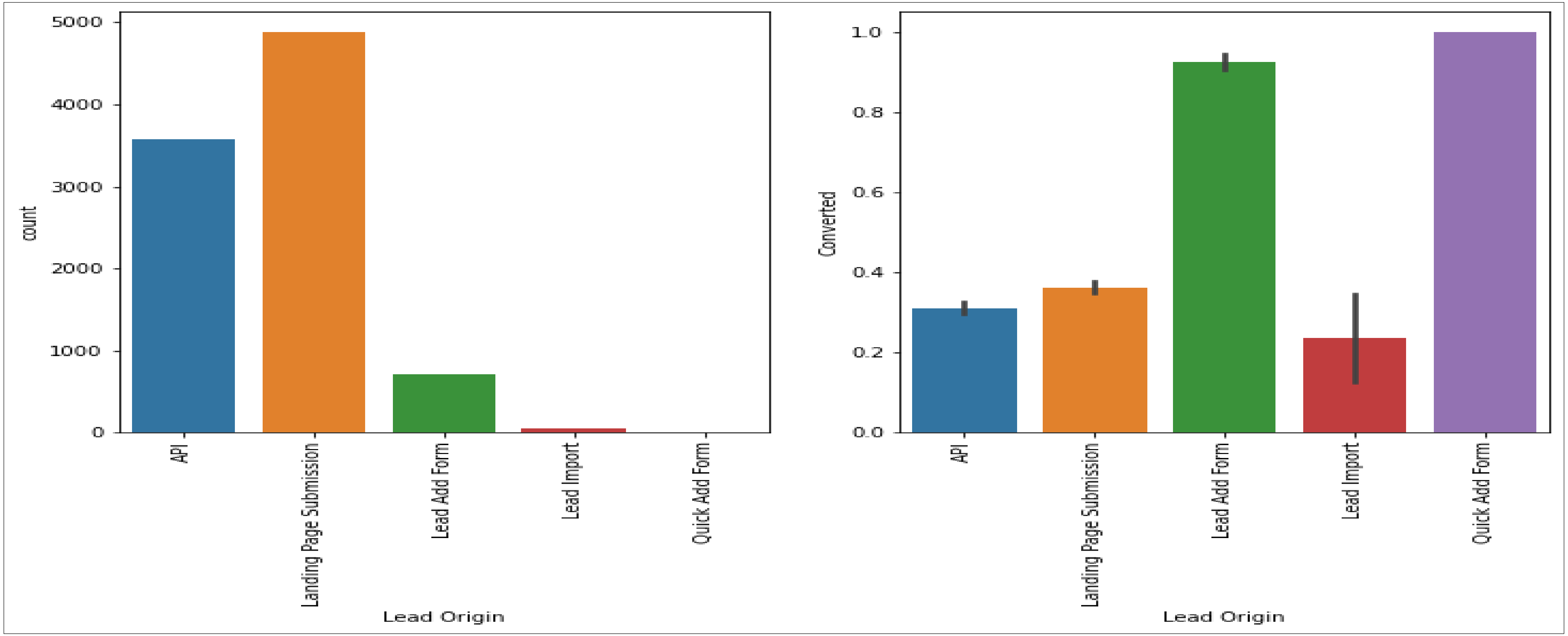
An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

- **What We need to Do:**
  - X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
  - The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

OVERALL APPROACH

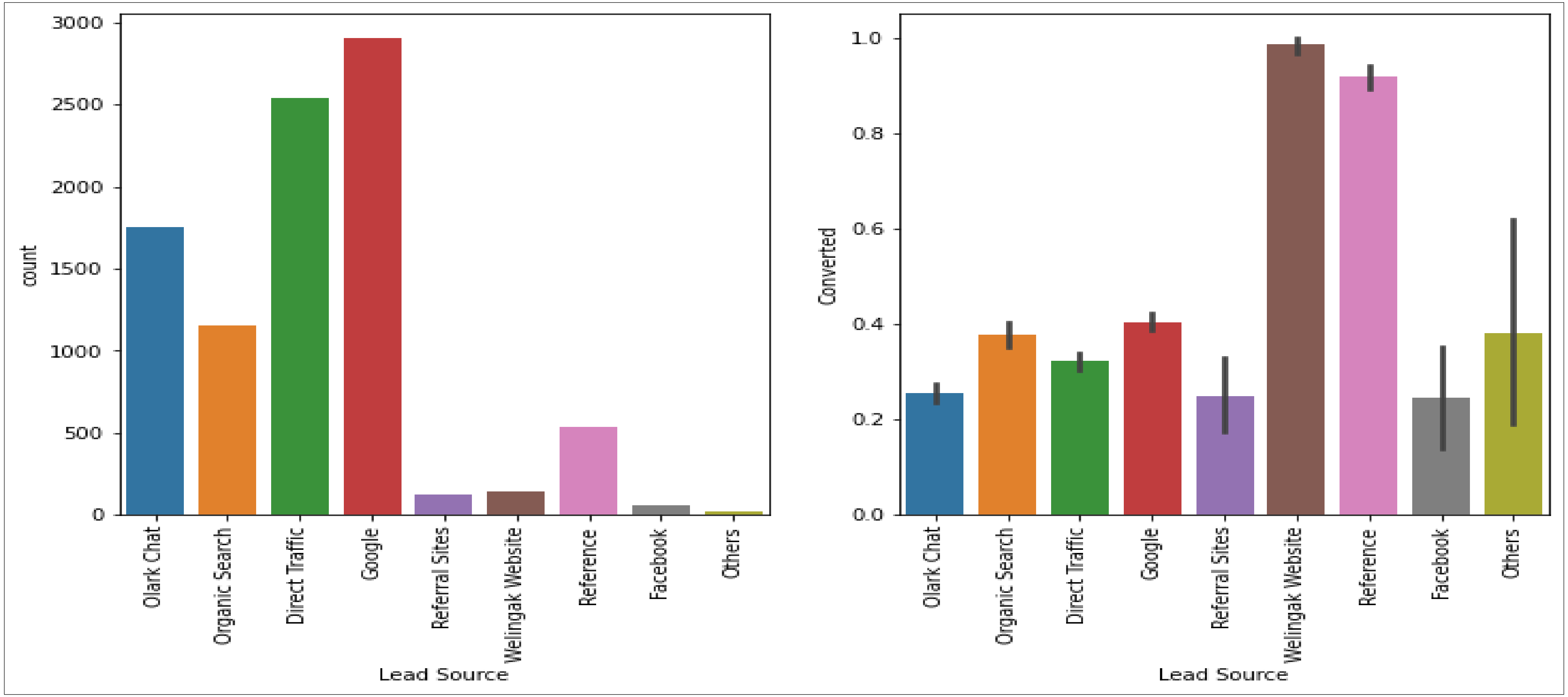
- Loading and Basic Checking of the Dataset
- Data Cleaning and Processing
  - Missing Value Treatment
  - Outliers Treatment
  - Checking the Columns with unique entries and Treatment accordingly
  - Combing Features to create more concise features etc.
- Univariate, Bivariate and Multivariate Analysis of the variables
- Data Preparation for Modelling (Creating Dummies, Train-Test Split etc.)
- Logistic Regression Model Building
  - Automatic Feature Selection and Manual Feature Selection using VIF and P Values
  - ROC and Threshold selection
  - Different Metrics Estimation for Train and Test data
  - Lead Score calculation for the whole data.
  - Important Feature Selection

EXPLORATORY DATA ANALYSIS



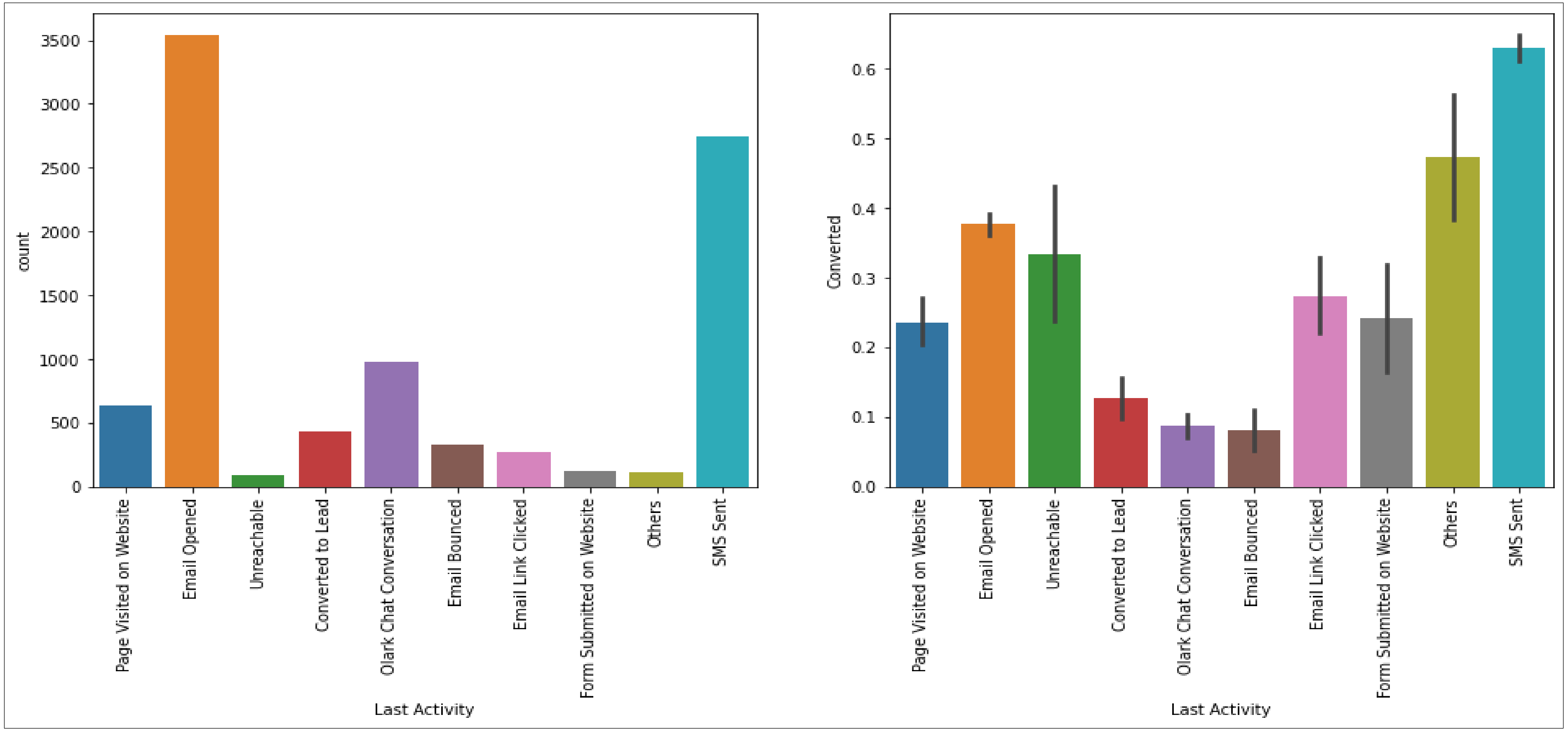
- **Lead Origin:** Maximum entry is from Landing Page Submission however; conversion rate is very high for Lead Add Form. It is almost 90%.

EXPLORATORY DATA ANALYSIS (CONT..)



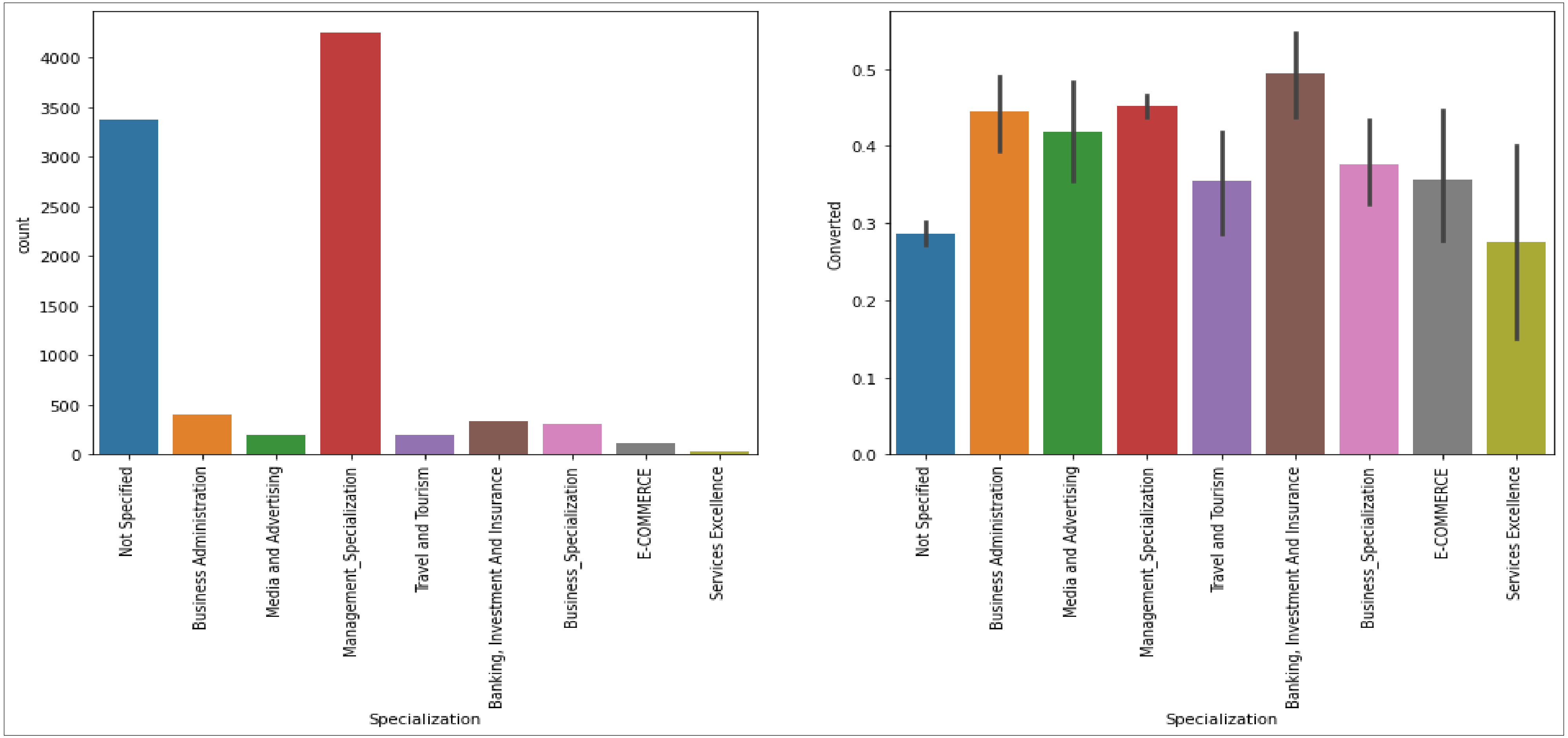
- **Lead Source:** Maximum Lead source is from Google. However, the conversion rate of Google is 40 %. For Reference and Welingak website the conversion rate is very high.

EXPLORATORY DATA ANALYSIS (CONT..)



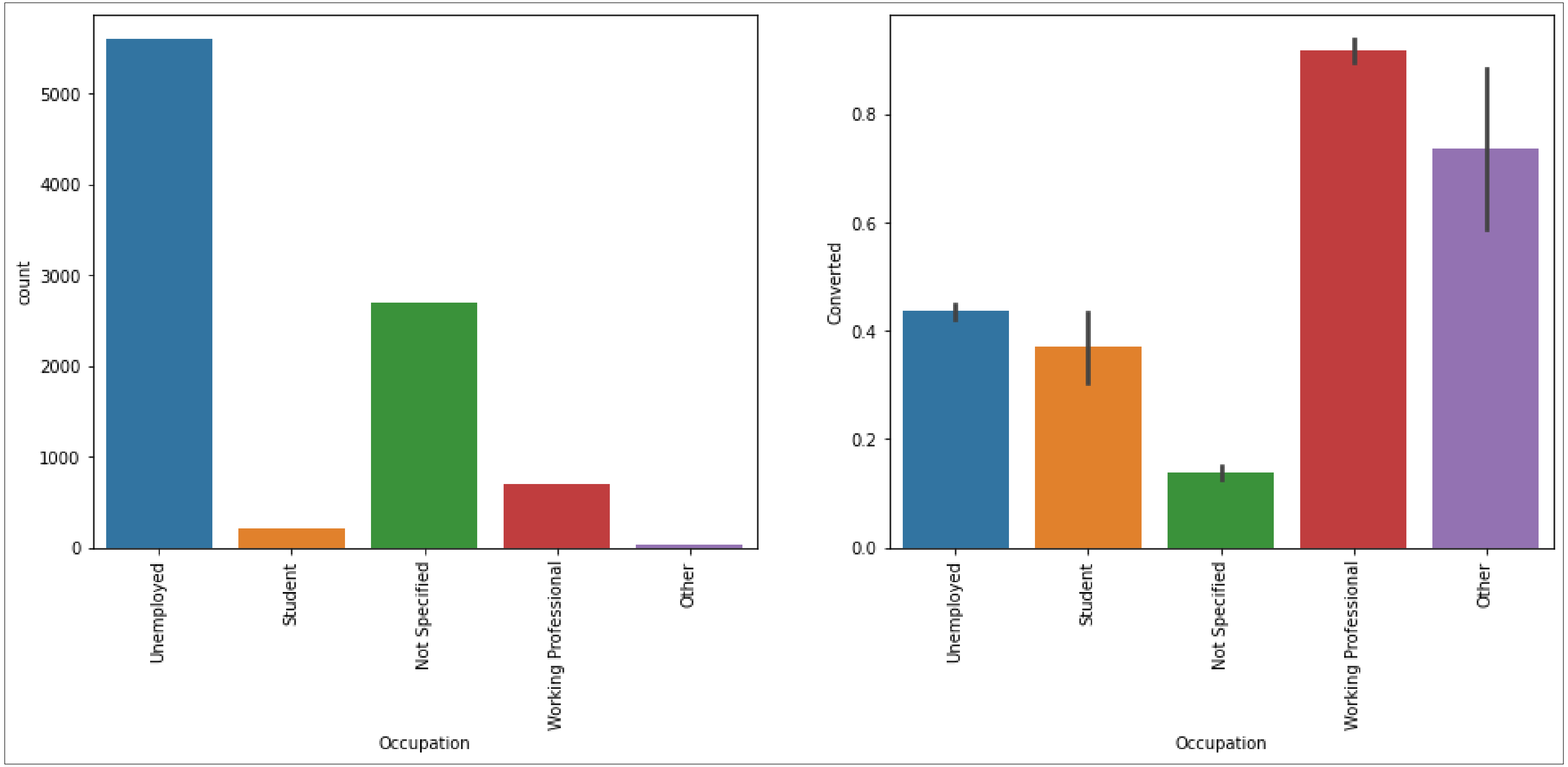
- **Last Activity:** For most of the leads last activity is Email Opened. However, the conversion rate is maximum for the leads whose last activity is SMS Sent.

EXPLORATORY DATA ANALYSIS (CONT..)



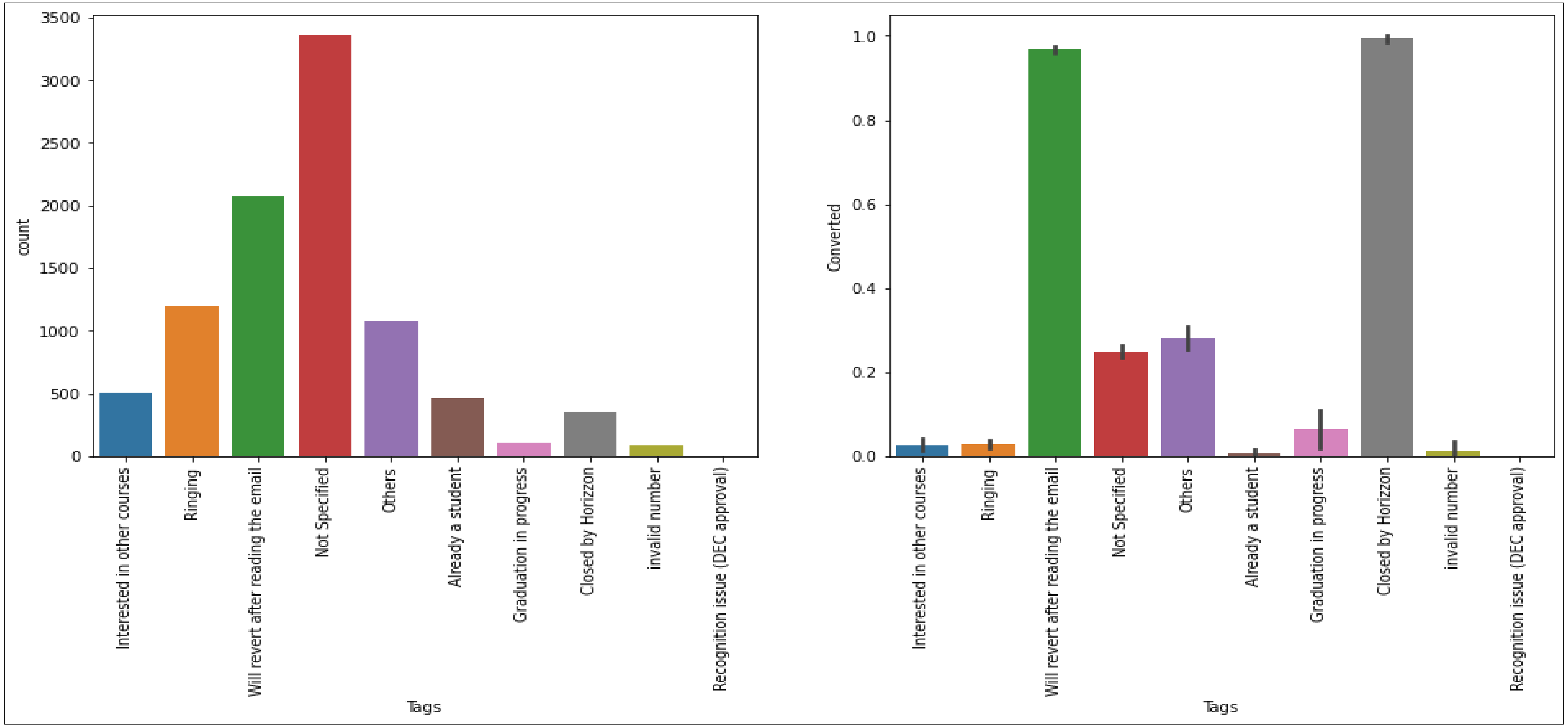
- **Specialization:** Maximum has mentioned Management as Specialization. Conversion rate for this specialization is also very good (45%). Maximum conversion rate is for Banking, Investment and Insurance Specialization.

EXPLORATORY DATA ANALYSIS (CONT..)



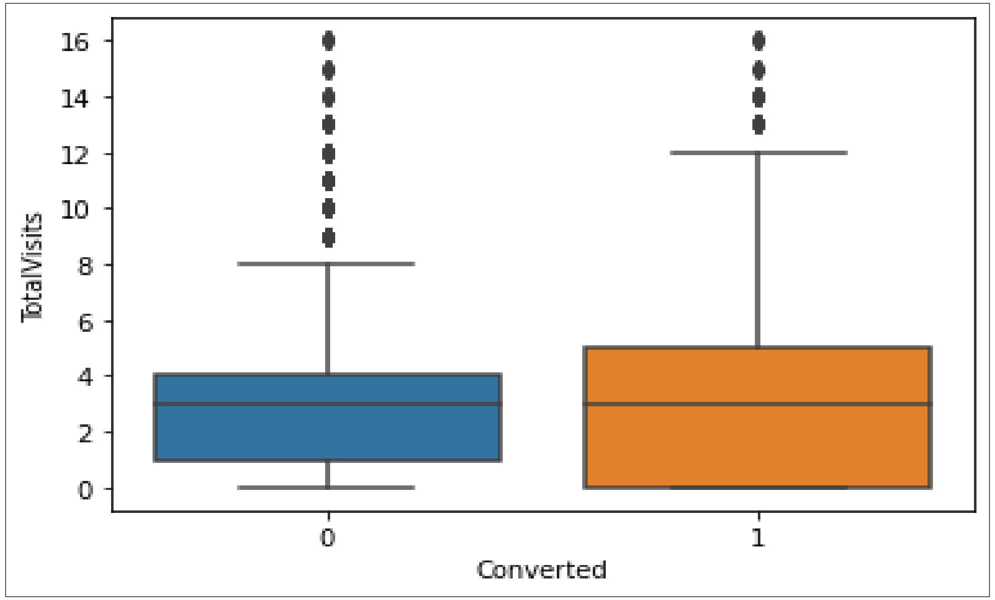
- **Occupation:** Maximum has mentioned Unemployed as Occupation. However, Conversion rate is maximum for working professional.

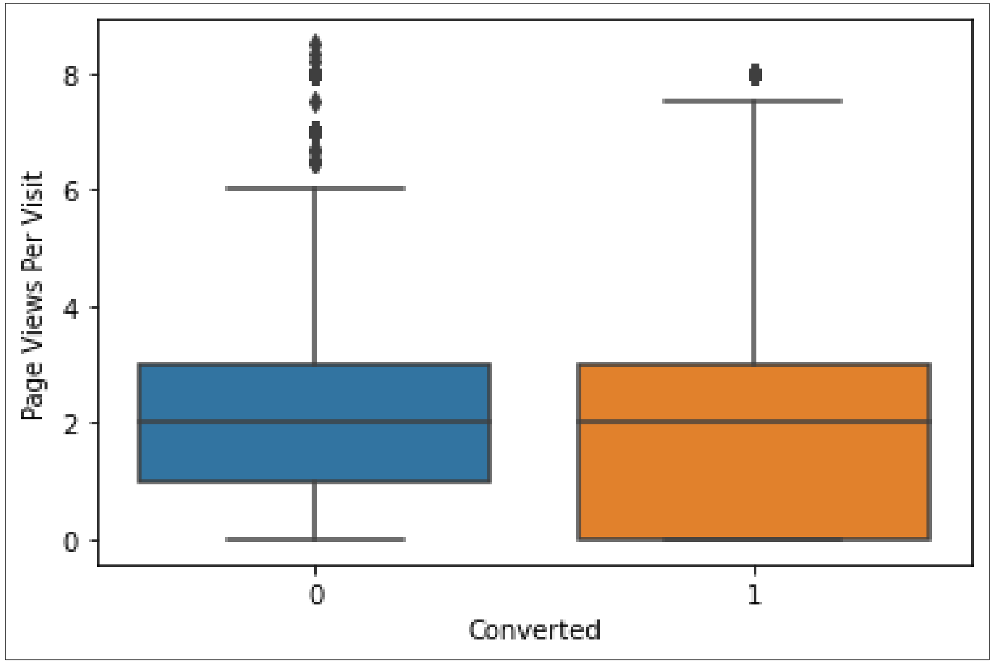
EXPLORATORY DATA ANALYSIS (CONT..)

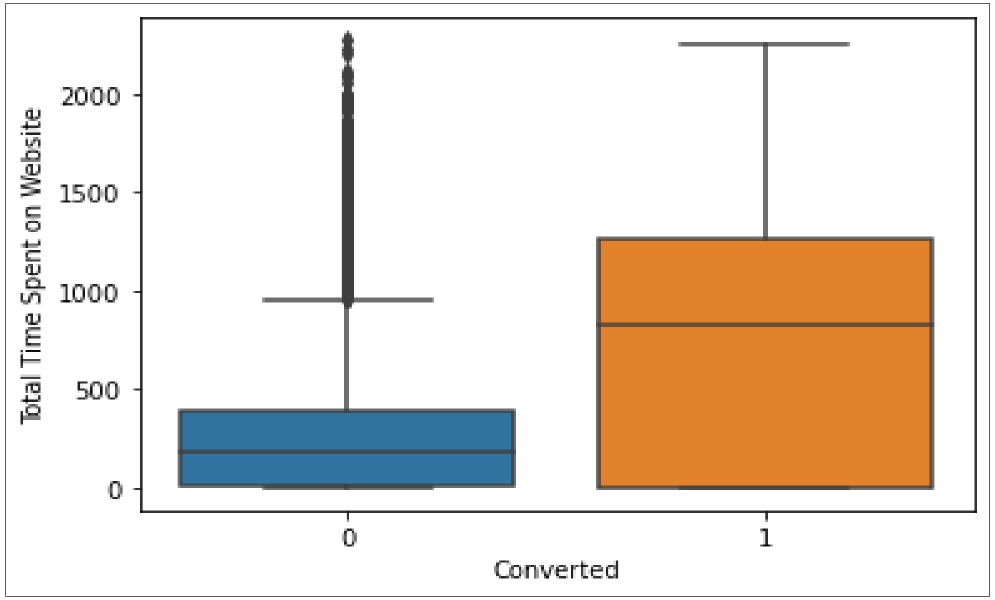


- **Tags:** Will revert after reading the email has good numbers as entry and conversion rate as well. For maximum case Tags has not been specified. Also, for Closed by Horizon has the highest conversion rate.

EXPLORATORY DATA ANALYSIS (CONT..)

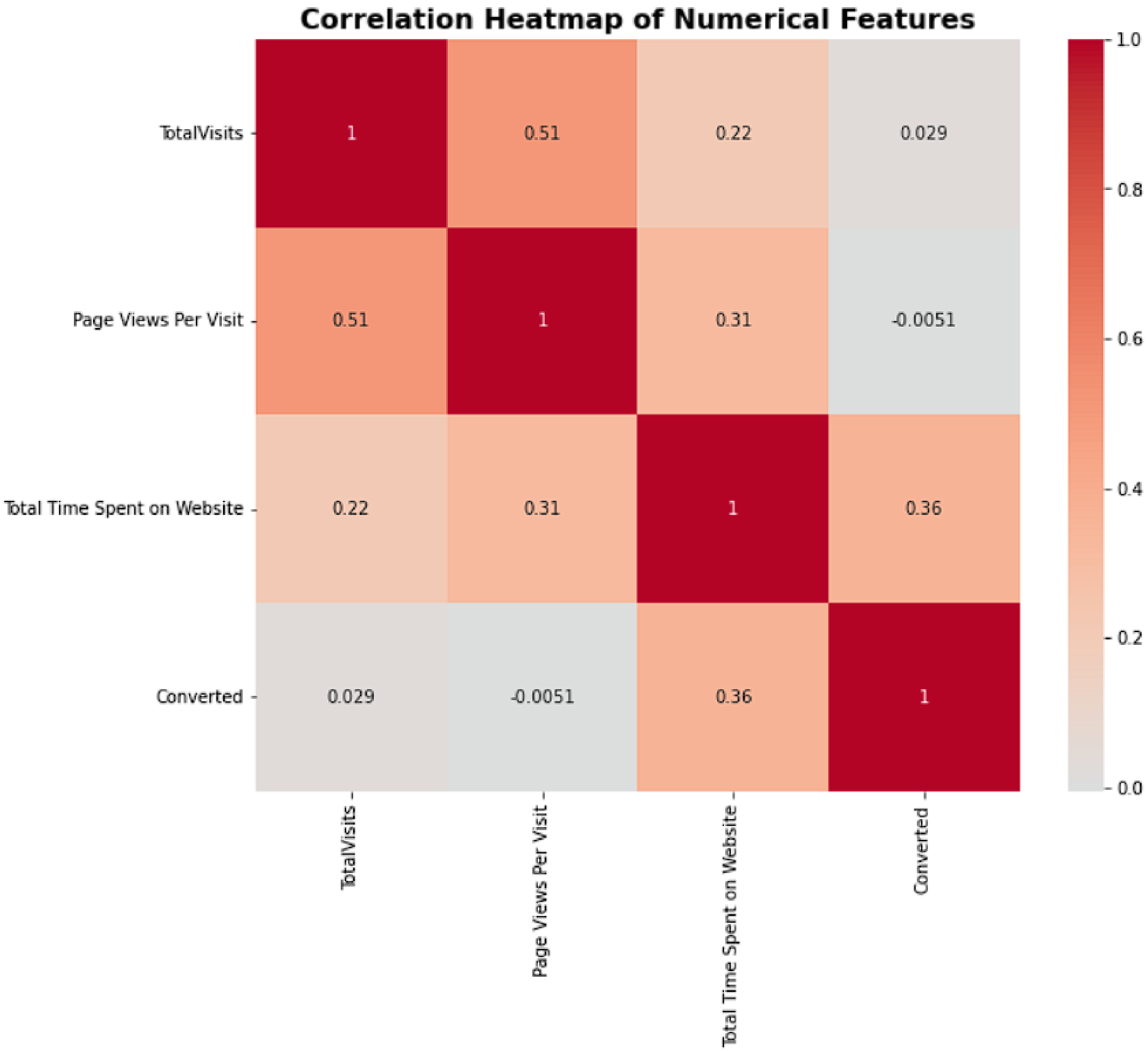






- **For Total Visits the median is almost at similar range for converted and non-converted.**
- **For Total time spent median is at higher range for converted.**
- **For Page views per visit median is at similar range.**

EXPLORATORY DATA ANALYSIS (CONT..)

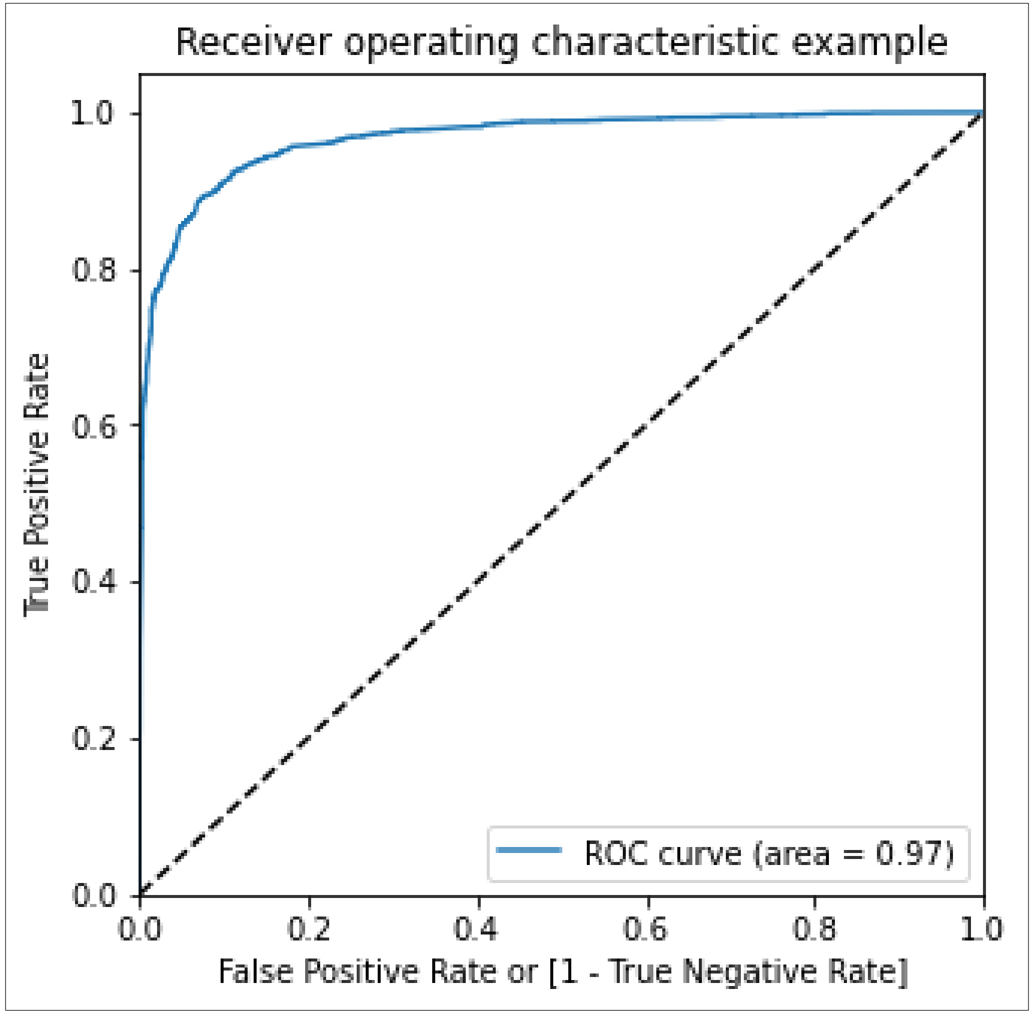


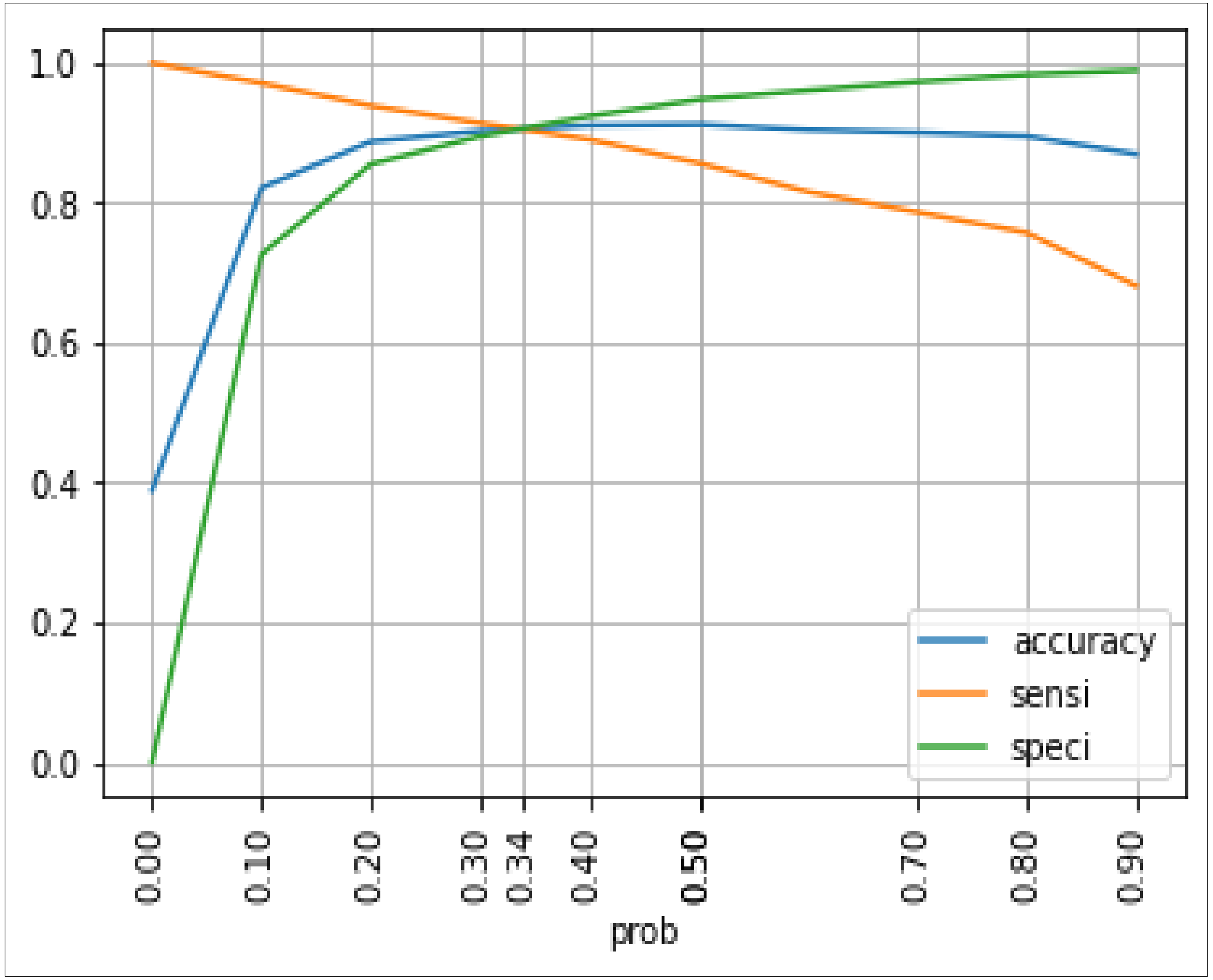
- From the above analysis it is evident that there is strong positive correlation of Converted with Total Time Spent on Website. There is positive but moderate correlation of converted with Total Visits. Page Views per visit is very loosely negatively correlated with Converted.

LOGISTIC REGRESSION MODELLING

- Dummy Variables were created for Categorical Variables.
- Train Test splitting of the dataset was carried out in 70:30 ratio
- Numerical Variables were scaled using Standard Scaler.
- Using RFE automatic feature selection were carried out. Thereafter, manually features were eliminated with P-value and VIF value.
- Using Final model predictions were carried out for train data with threshold value 0.5.
- Different metrices were computed to check the model efficiency.
- ROC curve and Threshold selection were carried out.
- Prediction on the test data were carried out and comparison of metrices were carried out.
- Lead Score was computed for the dataset.

LOGISTIC REGRESSION MODELLING (CONT..)





- ROC Curve area under the curve is 0.97.
- 0.34 was selected as threshold to be used to convert the probability value.

RESULTS & CONCLUSION

Metrices	Train Data	Test Data
Accuracy	0.90	0.91
Precision	0.86	0.86
Recall	0.90	0.92

Top 5 Impactful Features
-Tags_Closed by Horizzon - Tags_Already a student - Tags_Ringing - Tags_invalid number - Lead_Source_Welingak Website

Lead Score has been calculated and added to final database. Company X can refer to these and use the model to predict the Lead score for particular lead in future to convert them into hot leads.

Apart from that EDA also gives some idea on the parameters to be focused while targeting a Lead for further conversion